Linguistic DNA: investigating conceptual change in early modern English discourse

Article (Published Version)

Studia Neophilologica

ISSN: 0039-3274 (Print) 1651-2308 (Online) Journal homepage: http://www.tandfonline.com/loi/snec20

Linguistic DNA: Investigating Conceptual Change in Early Modern English Discourse

Susan Fitzmaurice, Justyna A. Robinson, Marc Alexander, Iona C. Hine, Seth Mehl & Fraser Dallachy

To cite this article: Susan Fitzmaurice, Justyna A. Robinson, Marc Alexander, Iona C. Hine, Seth Mehl & Fraser Dallachy (2017): Linguistic DNA: Investigating Conceptual Change in Early Modern English Discourse, Studia Neophilologica, DOI: 10.1080/00393274.2017.1333891

To link to this article: http://dx.doi.org/10.1080/00393274.2017.1333891

Published online: 14 Jun 2017.

STUDIA NEOPHILOLOGICA, 2017
https://doi.org/10.1080/00393274.2017.1333891

Routledge
Taylor & Francis Group

Linguistic DNA: Investigating Conceptual Change in Early Modern English Discourse

Susan Fitzmaurice, Justyna A. Robinson, Marc Alexander, Iona C. Hine, Seth Mehl and Fraser Dallachy

ABSTRACT

This article describes the background and premises of the AHRC-funded project, 'The Linguistic DNA of Modern Western Thought'. We offer an empirical, encyclopaedic approach to historical semantics regarding 'conceptual history', i.e. the history of concepts that shape thought, culture and society in a particular period. We relate the project to traditional work in conceptual and semantic history and define our object of study as the *discursive concept*, a category of meaning encoded linguistically as a cluster of expressions that co-occur in discourse. We describe our principal data source, EEBO-TCP, and introduce our key research interests, namely, the contexts of conceptual change, the semantic structure of lexical fields and the nature of lexicalisation pressure. We outline our computational processes, which build upon the theoretical definition of *discursive concepts*, to discover the linguistically encoded forms underpinning the discursive concepts we seek to identify in EEBO-TCP. Finally, we share preliminary results via a worked example, exploring the discursive contexts in which paradigmatic terms of key cultural concepts emerge. We consider the extent to which particular genres, discourses and users in the early modern period make paradigms, and examine the extent to which these contexts determine the characteristics of key concepts.

1. Introduction

Linguistic DNA is a three-year AHRC-funded collaborative research project in historical semantics and conceptual change in early modern English discourse.[1] The overarching objective is to discover relationships between words and ideas that exceed human intuition, with the help of computational methods for analysing big data.

This article lays out the background and premises of the project and its principal aims and themes. Discussion of the technical and theoretical challenges posed for the project forms the centre of this paper. Our aim is to consider the manner in which our goals and methods challenge current approaches to semantic enquiry and to offer an account of an empirical, encyclopaedic approach to historical semantics in relation to

CONTACT Susan Fitzmaurice s.fitzmaurice@sheffield.ac.uk University of Sheffield, United Kingdom.

[1]The project is based at the University of Sheffield (PI Susan Fitzmaurice, Co-I Michael Pidd, Matthew Groves, Iona Hine, Seth Mehl) with collaborating institutions University of Sussex (Co-I Justyna Robinson) and University of Glasgow (Co-I Marc Alexander, Fraser Dallachy, Brian Aitken). AHRC AH/M00614X/1.

what might be termed 'conceptual history', i.e. the history of concepts or paradigms that shape thought, culture and society in a particular period.

First, we contextualise the project in relation to traditional work in conceptual and semantic history, highlighting the innovations forwarded by Linguistic DNA. Then, we define our object of study as the *discursive concept*, a category of meaning which is encoded linguistically as a cluster of expressions that co-occur in discourse. It is encyclopaedic in data terms and as a set of expressions, it is not co-terminous with the keyword. We then describe our data, EEBO-TCP, and summarise the project's three principal research interests, which are built on the expertise of the project's primary and co-investigators, as well as on the nature of semantic and conceptual research. Next, we outline our computational processes, which build upon the theoretical definition of *discursive concepts*, to discover the linguistically encoded forms that underpin the discursive concepts we seek to identify in EEBO-TCP. Finally, we share preliminary results in the form of a worked example.

2. Background

The early modern period is arguably key to understanding the emergence of modern culture, historical thought and conceptual history in Europe. Some of the most influential accounts of the history of major concepts or paradigms rest upon comprehensive (but by no means exhaustive) surveys of canonical intellectual, philosophical and literary texts of the period (e.g. Porter 2001). The practice of this kind of deep reading through different theoretical lenses has yielded a set of paradigmatic terms felt to mark a watershed, the historical transition to modernity.

Discussions along these lines are typified by J.G.A. Pocock's (1972) treatment of English historical thought, and Reinhart Koselleck's (1998, 2004) on the German practice of *Begriffsgeschichte* (Conceptual History). Intellectual history, in these cases, has been constructed as change in the use (and meaning) of (paradigmatic) terms.[2] Such scholars examine the (synchronic) semantic complexity of given terms at particular moments and their (diachronic) complexity as historical, social and cultural keywords, as exemplified in Raymond Williams' *Keywords* (1983). Koselleck's method of *Begriffsgeschichte* (1998) explores the different societal and legal constructions of a concept such as 'marriage' in different periods (from the medieval period to the nineteenth century) and circumstances (peasant vs. 'Sub-peasant' marriages),[3] while Quentin Skinner (e.g. 1978) takes a rhetorical approach to conceptual change in the notion of the *State*. Pocock focuses on the study of political discourses, languages and metalanguage, attending to texts in their contexts. Thus his monumental *Barbarism and Religion* (1999–2015) uses Gibbon's *Decline and Fall of the Roman Empire* (1776–1789)

[2]For an overview of various approaches to the question, 'What is intellectual history?', see the contributions by Stefan Collini, Michael Biddiss, Quentin Skinner, J. A. G. Pocock and Bruce Kuklick in *History Today* (1985) http://www.historytoday.com/archive/history-today/volume-35-issue-10-october-1985

[3]Typically, Koselleck's (semasiological) approach starts with the term and then proceeds with investigating the nature of the changes in meaning that attend its use through time. He notes: 'From the perspective of linguistic history, the development of concepts within the new legal code and Romantic liberalism assumed the character of events. They then had repercussions for the entire linguistic structure within which marriages could be understood. It was not the diachronically given language as a whole that had changed, but rather its semantics and the new linguistic practices released thereby' (1998: 34).

as a basis for navigating the intellectual transformation of eighteenth-century Europe through the languages of Enlightenment law, religion and history.

What all these intellectual historians have in common is that they have produced accounts of the emergence and historical importance of such ideas for the history of Western political thought through the careful reading of (mostly) major historical texts. Their method thus consists of seeking the origins and development of these ideas by reconstructing them within their intellectual, political and historical contexts. Nevertheless, however intellectually ambitious this programme of historical inquiry may be, it suffers from the inevitable analytical and empirical limitations inherent in focusing on particular terms, selected from the vantage point of the historian.

This approach – restricting scrutiny to selected 'keywords', chosen on a subjective basis – more clearly typifies work in historical semantics and historical lexicography within the field of linguistics (Fitzmaurice 2016). Scholars have tended to produce (semasiological) studies of the changes over time in the meanings of individual words, for example, *enthusiasm* (Tucker 1972), and *story* or *evidence* (Wierzbicka 2010). Historical semantics and dialectology involve (onomasiological) studies of the changing lexical expression of concepts such as TRUTH (Lenker 2007) or ANGER (Geeraerts *et al.* 2012), or of the varying lexical realisation of more concrete concepts such as LEGGINGS in the Dutch dialects of the Netherlands and Belgium (Geeraerts 1997). More generally, corpus semantics studies rely upon using as a starting point the individual lexical item or a lexical field in order to investigate meaning change or change in the structure of a lexical field over time.

Semasiological and onomasiological approaches have been central therefore to the study of word meaning and, by extension, conceptual meaning, in the history of ideas. Our project builds on the insights developed, but goes much further. We are intrigued by the possibility of discovering relationships among words and ideas that we cannot intuit in a universe of early modern English printed discourse. In other words, we are interested in identifying, in a much more thorough and robust way, what emerging and important concepts writers produced in their own time in their own writing – and how those concepts were constellated.

Until now, it has not been possible to apply a bottom-up approach to the investigation of conceptual structure, but new developments in corpus studies offer an exciting opportunity to undertake just such a project. The Linguistic DNA team is currently harnessing an array of computational methods together with close reading in order to model concepts in early modern English discourse.

To this end, we have developed a set of key questions about semantic and conceptual change in Early Modern English:

(1) What are the characteristics of a paradigmatic term in early modern English discourse?
(2) What is the nature of the lexical complexity of historical concepts?
(3) What conceptual fields undergo rapid or unusual increases in the volume of terms used?

Together these questions guide our research into three parallel themes that lie at the heart of the project: (1) the contexts of conceptual change, (2) the semantic structure of

lexical fields, and (3) the nature of lexicalisation pressure. The rest of this paper sets out how we are tackling these questions. Following a brief discussion about the epistemology underpinning the project, we map out the concerns of these three research areas. We then describe our methodology and report some indicative results and directions for continuing research. Our aim is to compare what we discover in the process of our distant corpus reading using computational methods, with what earlier scholars of intellectual history intuited through the careful close reading of canonical texts.

3. Epistemology: What is a concept?

The challenge of investigating what people (in particular periods) would have considered to be emerging and important cultural and political concepts in their own time involves careful searching of the literature of that time to see what emerges as important. This task cannot be undertaken by identifying a set of concepts in advance and mining the literature of the period to ascertain the impact made by those concepts. Our approach is neither semasiological, whereby we track the progress and historical fortunes of a particular term, such as *marriage, democracy* or *evidence*, nor is it onomasiological, whereby we inspect the paradigmatic content of a more abstract, yet given, notion such as TRUTH or ANGER. We have to take a further step back, to consider the kind of analysis that precedes the implementation of either a semasiological or an onomasiological study of the lexical material we might construct as a concept (e.g. as indicated by a keyword).

The theoretical work of defining conceptual structure involves finding a model of semantics in which meaning is systematically constructed as language in context. For Linguistic DNA, a concept is not assumed to be the same thing as a keyword; accordingly, it is not coterminous with a keyword. Thus a concept is not encoded as a lexical word or as a set of words in traditional semantic relationships such as polysemy, synonymy (or onomasiological alternation, as in Geeraerts et al. 2012, Geeraerts 1997, Geeraerts et al. 1994, and additional work undertaken by the Quantitative Lexicology and Variational Linguistics research group), or hyponymy (as in the Historical Thesaurus of English, WordNet, or typical approaches to thesauri). Instead, as we construct a concept as greater than and/or other than the lexical items it contains (and their semantic relations), it is necessary to look beyond the level of the word to discover the broader meaning relationships that the concept encapsulates, that is, into discourse.

In a nontrivial fashion, seeing concepts in discourse might be constructed as a dynamic process in which the 'reader' finds or constructs meaning at different levels, from broad to increasingly narrow domains. Thus, the human analysis of context starts in material, historical, cultural, economic and political worlds, subsequently focusing on contexts where interactional norms and goals obtain, thence to discursive domains and ultimately to the utterances themselves. The further away from our own present-day world the text is, the more necessary it is to bring encyclopaedic knowledge to its analysis. In this understanding, close reading is a method of human analysis which is systematic, rigorous, principled, critical, creative and qualitative, each level of meaning informing the next. Indeed, the balance of quantitative corpus methods and deep qualitative discourse analytic methods (cf. Taavitsainen and Fitzmaurice, 2007: 15–20; Jucker and Taavitsainen 2013) is central to Linguistic DNA.

The consequence of apprehending a text in the dynamic way outlined is the systematic construction of meaning as language in context. Meaning, thus defined, consists of constant meaning (from the word) and contingent meaning (from the context), or of both semantics and pragmatics, a combination that allows the concept to be constructed from linguistically encoded form.[4] Crucially, the difference between constant and contingent meaning may be tenuous and arbitrary and it is not categorical. Accordingly, we assume that meanings are created systemically by virtue of their contexts of use, from material, historical, social and cultural, etc. via discoursal, to interactional and utterance contexts. Thus meanings can encompass the domains of discourse, pragmatics, semantics and topics. We take the view that the word is not coterminous with the concept. The core semantic construct for Linguistic DNA is therefore the *discursive concept*, the opposite construct of the query term that is central to semantics, specifically, corpus semantics.

Discursive concepts have the following attributes: they are not unified, discrete or compositional meanings. Instead, they are ontological constructs, and as we will demonstrate, in data terms, they are equivalent to encyclopaedic entries. Specifically, the linguistically encoded form that underpins the discursive concept is the co-occurrence cluster. In any particular historical moment, a concept might not be encapsulated in any single word, phrase or construction; instead it will be observable only via a complete set of words, phrases or constructions in syntagmatic or paradigmatic relations to each other in discourse. We therefore operationalise the concept at a supralexical level; it will be traceable in the associative relations among words distributed in texts. Indeed, lexicalisation might be the final stage in the linguistic realisation of a concept.[5]

Humans can 'see' and thus analyse meaning in discursive contexts; we need to consider what computational processes can be used to automate this kind of meaning construction. The discursive construction of conceptual structure can be approached using the notion of encyclopaedic meaning (Evans 2015). The kind of co-occurrence marking the relationships among the words that encode the discursive concept belongs to computational distributional semantics. Co-occurrence in this model captures association: a notion of relatedness that is much looser than that captured in formal synonymy (cf. Heylen et al. 2008) or strict collocation (cf. Manning & Schütze 2001, Chapter 5). Thus the discursive relations that might obtain among words distributed across a text – *news*, *politics*, *election*, *Clinton*, *polls* – are different from the relations of synonymy (*stranger*, *foreigner*; or *abode*, *residence*), hyponymy (*emotion – anger*), or the analogic relations (*king*: *queen*, *man*: *woman*) among sets of expressions.[6]

4. Data: The universe of early modern English printed discourse

As a project, we are interested in finding an objective way to identify concepts and to observe how they form and change in a particular universe of English printed discourse

[4]The theory of lexical meaning underpinning this view is derived in part from cognitive semantics (cf. Evans, 2015; 2009). Most approaches to semantics utilise context, whether this is the relatively local (e.g. utterance; sentence) or the broadest kinds of context, as instanced by the social and cognitive worlds that underpin many corpus linguistic approaches to semantics, such as that of QLVL.

[5]Note too the relevance for Linguistic DNA of Lehrer's (1992) work on lexical field theory, Fillmore's encyclopaedic view of semantic organisation and Lakoff's notion of frames (Geeraerts, 2010).

[6]The similarity among expressions related by synonymy, hyponymy or analogy might be investigated using vector space analysis (Heylen *et al.* 2008; Turney and Pantel 2010).

in the early modern period (1500–1800). Our primary source of data is *Early English Books Online* as curated by the Text Creation Partnership (henceforth EEBO-TCP).

EEBO-TCP comprises more than 50 000 document transcriptions produced from the pre-digital commercial microfilm collection *Early English Books*. A commercial product at its outset, *Early English Books* used the Short Title Catalogues to identify items printed in a British language and/or in the British Isles prior to 1700, microfilming copies largely from the collection of major libraries (such as the British Museum) with a preference for first editions. With the advent of the digital age, many of these images were scanned and made available to subscribers over the internet as *Early English Books Online*. In addition, a partnership drawing funding and direction from academic institutions including the Universities of Michigan and Oxford undertook full-key transcription of an expanding subset of this data. This endeavour, under the auspices of the Text Creation Partnership, is responsible for EEBO-TCP and its smaller siblings: EVANS-TCP and ECCO-TCP (representing early American printing and a small subset of Eighteenth Century Collections Online respectively). The latter pair and the first phase of EEBO-TCP (amounting to about 25 000 pre-1700 items) are now in the public domain, while academic institutions in the UK have access to the full set of transcriptions through Jisc.[7]

In the grandest terms, Linguistic DNA can therefore claim to access and analyse the 'universe of early modern English print'. Yet we are compelled to be cautious in certain respects. The composition of EEBO is haphazard by nature: it reflects the portion of printed matter that survived to be catalogued and microfilmed. Book historians remind us that there are patterns in what survives: large reference works were more likely to remain secure (and perhaps unread) in libraries, whereas ephemera seldom survive unless someone cared to collect them. Some genres (recipe books, grammars) were eminently disposable as they wore out, or were supplanted by 'new improved' versions. The nature of the texts represented in EEBO-TCP is therefore accidentally but in some very particular ways unrepresentative of what was printed (cf. Bruni & Pettegree 2016). In corpus linguistic terms, it is also undesigned: this is not a corpus built through intentional sampling to offer a representative perspective of early Modern English discourse. It is a digital collection, not a corpus. Importantly, it is a collection skewed toward some particular interests, insofar as the subset of EEBO that has been transcribed consists of items chosen because of a perceived historical, literary or perhaps book historical interest. These facts about our principal dataset are something to be aware of, and these are limitations we accept. (And of course the discourse of early modern England, like EEBO itself, was not limited to English-language texts.)

In addition to not being a corpus, by default EEBO-TCP is not prepared for linguistic analysis. Its native annotation, TEI encoding, records characteristics of print: mise-en-page features such as change in typeface, divisions of text and decorated initials. It strives to be utterly faithful to the printed page, so that where a typesetter has misplaced a letter, the resulting mis-constructed word constitutes the 'ideal' transcription. Of course, in a search-and-find digital culture, this is quickly sub-optimal: many of EEBO-TCP's most avid users approach it with keyword search, and it was quickly necessary to provide some tools to allow for spelling variation. Nonetheless, the

[7]For a more detailed history, cf. Gadd (2009) and Kichuk (2007).

transcribed texts are not annotated for linguistic analysis. To facilitate that, and to permit calculations based on standardised spellings, lemmas and part-of-speech, Linguistic DNA employs the MorphAdorner pipeline, developed at Northwestern University by Martin Mueller and Philip Burns (Burns 2013).

MorphAdorner traces its origins to the EEBO interface and to work with classical Greek texts during the Wordhoard project. In practice, its trigram tagger uses a hidden Markov model and a variant of the Viterbi algorithm to assign part-of-speech and regularise spellings, drawing on a lexicon optimised for historical English. It outputs the original forms marked up with XML tags assigning a unique token identifier for each item together with lemma, part-of-speech and standard spelling. When tested on sample texts, MorphAdorner 2.0 made more accurate decisions than the main alternative (VARD), especially with the earliest material, leading to its incorporation into Linguistic DNA's preparatory steps. Its part-of-speech tagset (NUPOS) is idiosyncratic yet functional for our purposes.[8]

5. Research areas

5.1. *Contexts of concepts and conceptual change*

One of the principal interests of the project is the contexts in which concepts emerge and change in early English discourses. These contexts include the narrowly textual and co-textual domains with which historical discourse analysis and historical semantics are usually concerned. Yet we are interested also in other kinds of context: the social, historical and literary dimensions documented in other disciplines.

To this end, when analysing quantitative data, we seek to link the interpretation of our data tightly with examples from particular texts and passages of text, while comparing these with the work of other scholarly readers, including social historians, book historians, and literary scholars. This work is thus deeply concerned with the connection between distant and close methods of reading, and the interplay between the mechanised documentation of language information and human understanding of it. What is necessary to know in order to appreciate the complex evolution of communication attested by this body of English print? Consider the example of John Speed's two-volume *History of Great Britaine*, published in 1611 (discussed below §7). Knowing that Great Britain was itself a novelty, conceived upon James VI's inheritance of the English throne (1603) – or as Schama (2001) would have it, in James' head – leads us to attend to the rhetorical and political dimensions of Speed's work. Armed with this knowledge we can appreciate how his lexical choices are part of the author's endeavour to present Britain as one cultured nation. These lexical choices include the use of relatively new words or meanings derived from prestigious settings such as classical texts.

We seek to re-examine the historical backdrop to conceptual change. We use Linguistic DNA data to interrogate existing accounts of paradigm shifts, taking the view that the supra-canonical set of texts present in EEBO-TCP can be harnessed to question and probe canonical ideas about historical conceptual change. We look to see

[8]For further information, see: morphadorner.northwestern.edu.

if there are patterns in the emergence of new ideas, exemplified by particular publications or textual contexts. At the same time, we take responsibility for asking epistemological questions about the use of linguistic evidence to trace historical discourses. Advances in our understanding of the contexts of conceptual change is critical to the wider endeavour of Linguistic DNA to discover how and where movements in the sixteenth, seventeenth and possibly eighteenth centuries affect the shape not only of early Modern English language but the linguistic modes of thought that continue to structure society today.

5.2 *The semantic structure of lexical fields*

We are also interested in linguistic semantics and traditional semantic relations as they relate to the computational processes developed in the project (see section 6 below), and lexis in historical texts. We aim to assess the ways that the analyses generated by the Linguistic DNA processes, which include particular measures of the co-occurrence of expressions,[9] might correlate with elements of meaning, such as degrees of polysemy or synonymy, or degrees of constant and contingent meaning, for a given lemma. Therefore we approach our research both semasiologically and onomasiologically, looking at the various meanings projected by groups of co-occurring words, and the various constellations of words for expressing a given meaning. So we are particularly interested in how particular lexical co-occurrence patterns (identified by our processes and measured in statistical terms of proximity and strength of association) can be construed in terms of traditional semantic relations such as polysemy. For example, our processes may identify lexical pairs which demonstrate a strong attraction (e.g. *virgin, marriage*) which expands into a trio via the strong association with items such as *Christ*, yielding a discursive concept centring on Christian celibacy. At the same time, the same pair might expand into a set of trios via the pair's strong association with *maiden, youth* or *matrimony*, arguably suggesting another discursive concept. Can such relationships between such co-occurrence clusters, as identified by our computational processes, inform our understanding of polysemy, or of other traditional semantic relations?

Further, we examine the possibility of distinguishing between *lexical polysemy* and what we call *discursive polysemy* in historical texts: the former indicates relatively discrete constant meanings of a word that cannot be evoked simultaneously, while the latter indicates broadly discursive conceptual variation related to textual contexts beyond the level of the utterance.

To address these questions, we further analyse Linguistic DNA process outputs, employing statistical methods, comparing findings to the OED and HTE, and exploring EEBO texts via close reading.

5.3. *Lexicalisation pressure*

Finally, we are interested in the interaction between traditional thesaurus-style semantic categories and the discursive concepts which are being identified by

[9]These include measures such as Pointwise Mutual Information (PMI) distribution for a given lemma, or the number of lexical trios built up from a given lexical pair.

Linguistic DNA. In particular, we look for areas of vocabulary which have undergone rapid expansion or contraction, and seek patterns in these changes, especially whether the emergence of new words appears to drive or be driven by pressure from the discursive concept sets.

This work uses the *Historical Thesaurus of English* (Kay et al. 2016) as a dataset for lexicographical semantic categories. The *Historical Thesaurus* draws on the contents of the 2nd edition of the *Oxford English Dictionary* and its supplements along with extra material from sources such as *A Thesaurus of Old English* (Roberts & Kay 1995). It separates the different senses of words contained in these sources and categorises each of them. In this way, over 800 000 words are placed into around 225 000 meaning-based categories, designated a posteriori by the compilers as best reflecting the distinctions in meaning they perceived in the data. The categories are hierarchically arranged so that the further down the hierarchy tree a user explores, the more detailed and precise the distinctions of meaning become. The word senses contained in the *Thesaurus* are arranged within categories by the date at which they first came into use, from the Old English period onwards.

As the *Thesaurus* contains detailed dating information for its word senses, the size of a category (i.e. the number of active words it contains) can be graphed across time. Most categories grow as time proceeds, and follow a similar change in size over the course of the history of the language. Some categories, however, do not follow this standard trajectory and may display sudden dramatic expansions or equally precipitous declines. These categories are of most interest. Theories of linguistic evolution (e.g. Samuels 1972; Smith 1996), posit that changes in the vocabulary of a language are the result of both internal linguistic pressures and external historical pressures. The discursive associations generated by Linguistic DNA's processes should provide evidence for both kinds of pressure; for instance, particularly prevalent sets of words indicate culturally salient ideas whilst the breadth and shape of the associated vocabulary (including its part of speech distribution) constitutes a significant part of the linguistic context which helps shape a *Thesaurus* category's development.

Further to identifying the contexts which help to drive atypical change in the size of a category, we investigate the internal structure of these categories. Of particular interest is whether different parts of speech pertaining to the *Thesaurus* category expand at different rates. It may be that an expansion in one part of speech is particularly likely to catalyse creation of other parts, or indeed, there may be no relation between parts of speech at all so that, for example, adjectives belonging to a *Thesaurus* category experience bursts of expansion which are completely separate from their associated nouns, verbs or adverbs.

Pursuing these goals should provide a valuable perspective on the discursive associations which emerge from Linguistic DNA processes. As Linguistic DNA's discursive concepts emerge out of the computational marshalling of vast amounts of textual data, side-by-side analysis will shed light on the relationship between discourse-context meaning and human-categorised meaning.

6. Processes of concept modelling: Computational distributional semantics

In order to identify concepts by inspecting the language itself rather than starting with a selected term or keyword and investigating its characteristics, Linguistic DNA has

developed computational processes of indexing and analysing lexis in the texts. The raw frequency of occurrence of every single word in each text in the dataset, for single documents and sets of documents, is calculated and indexed. Each word index includes tokens, types and lemmas, allowing us to calibrate frequencies both within documents and across our data, taking into account chronological metadata when relevant.

The next step involves calculating the co-occurrence patterns, for example, of every noun in each text with every other noun in the data (i.e. word pairs) in different proximity windows. We started with the small window sizes (plus or minus one word in the range of a query term) typical of many corpus linguistics studies, noting that Heylen et al. (2008) concluded that semantic similarity occurs in small windows. However, the operationalization of a concept as a discursive construct and the hypothesis that these concepts manifest through associations of words that are visible at the discursive level require the investigation of co-occurrence patterns across larger windows of text. Accordingly, the processor tracks the co-occurrence patterns of words in windows of 20 words (10 to the left and right of each query word, W20), 100 words (50 to left and right, W100) and of 200 words (W200). These pairs are operationalized as symmetrical co-occurrence matrices in which each row and each column contain all lemmas in the data set or subset (cf. Turney and Pantel 2010).

In addition to using frequency information about the occurrence of each lemma and its co-occurrence with other lemmas in different proximity windows, Linguistic DNA calculates Pointwise Mutual Information (PMI), which measures word associations by comparing observed co-occurrences with what might be expected in a random distribution of the same lexical items. This measure enables us to distinguish between pairs of words that occur more than expected by chance (showing a strong, frequent association in the form of high positive PMI values); pairs of words that occur roughly as much as expected by chance (indicated by PMI values near zero); and pairs of words that occur less than expected by chance (marking a notably rare association in the form of low negative PMI values).[10] PMI measurements depend on measures of probability (cf. Fano 1961), and we adopt an innovative approach to linguistic probability for PMI. A probability is properly defined as a number of observed occurrences against a baseline number of possible occurrences (Sheskin 2004: 88). Unlike much previous work in corpus linguistics and NLP (cf. Turney and Pantel 2010), but in line with Wallis and Bowie (2012), we argue against measuring linguistic probabilities using a baseline of the total number of words in the texts. Such a measure results in an artificially low probability, because the total number of words in the texts is in fact far larger than the number of possible occurrences of the given observed lemma. That is, the given observed lemma cannot in fact alternate with all other lexical items in the text. We have therefore introduced an operationalisation of PMI based on Part of Speech, such that an observed noun, for example, is counted against a baseline of the total number of nouns, rather than the total number of words. This operationalisation eliminates a huge number of invariant Type C terms that cannot vary with the observed term, and which

[10]When lemmas co-occur as often as expected by chance, that does not mean that they co-occur by chance. Language is a tool for communication, which includes systematic constraints, and lexical items will only rarely, if ever, co-occur by chance.

therefore should not be included in a probabilistic baseline (Wallis & Bowie 2012).[11] While we acknowledge that this is not a perfect solution to the problem, we nonetheless posit that it is an improvement on traditional measures of PMI.

Linguistic DNA's process thus captures patterns of association based on frequency measures, proximity, co-occurrences, probability and density. These procedures are applied in an iterative fashion, to different segments of the data (by decade, etc.) in order to test the robustness of the processes. Each iterative analysis is performed on a small portion of the data and then assessed in an exploratory way via close reading. Hypotheses are formed and then tested against other subsets of data. A finalised set of computational processes will be run on the entire data set in 2017.

The next phase in the construction of the Linguistic DNA process involves a ground-breaking technique to move from pairs of co-occurring lemmas to trios and larger lexical sets. To undertake this task, we create an asymmetrical co-occurrence matrix in which rows represent observed co-occurring lemma pairs, and columns represent individual lemmas. This co-occurrence matrix identifies co-occurring trios of words within a window around a central node-word; PMI scores for these co-occurrences are calculated as well. Calculations for larger lexical sets can be built up iteratively in the same manner. These associative sets – co-occurrence clusters – are the raw linguistically encoded material for discursive concepts.

7. Some preliminary results

How do you demonstrate a discursive concept? A thorough answer to that question lies in the future of the Linguistic DNA project, but it is already possible to draw out something from preliminary data, even before we apply the part-of-speech-refined baselines described above (§6). In what follows, we illustrate the benefits of a discursive approach to conceptual modelling through an example focused on the node-word *valour*. As the focus suggests, this example does not adhere to Linguistic DNA's ideals of bottom-up data-driven concept detection, which involves querying the potential concept-role of any and every recurring lemma. The worked example below, which begins with a pre-selected lemma, should serve to illustrate our perspective on the work. While present output is limited to co-occurring pairs, probing the space around words where shifts in meaning might be expected has allowed us to test and inspect our operationalisation of the discursive space using different window sizes (as described above, §6). Here, we use Linguistic DNA outputs to test a finding first observed via different methods (Hine 2014).

Specifically, Hine selected the English word *valour*, a word introduced into the English Bible in 1611 and repeatedly used to translate a Hebrew concept with the phrase 'mighty … of valour'. A basic quantitative analysis of EEBO-TCP indicated an atypical increase in the frequency of *valour* and the related adjective *valiant*. This led Hine to suggest that 'the imputation of boldness and courage … came to the fore in the late sixteenth-century', connecting its growth with 'extremely high frequency in texts

[11]Precedent for using grammatically-oriented baselines to reduce invariant Type C terms is found in Bowie et al. (2013), who measure perfect constructions against a baseline of all perfect auxiliaries, and against a baseline of all verb phrases. Such grammatically-oriented baselines have not, to our knowledge, been implemented for PMI calculations.

translated from French, Latin, Spanish and Italian' as exemplified in *The historie of Guicciardin conteining the vvarres of Italie and other partes* (1559; EEBO-TCP A02329). She observed too that 'the words normally appear in certain genres: conduct books concerned with warfare and chivalric behaviour; and chronicles of past history' (2014: 163). Despite being based on a quantitative survey of frequency by decade, Hine's observations rely upon subsequent interpretive reading of texts and contexts to arrive at a conclusion about the factors driving semantic change. Is it possible to gain similar evidence of the 'fact' of semantic change by inspecting and comparing co-occurrence data for *valour* in chronologically-defined subsets of EEBO-TCP?

For the purposes of this example, Linguistic DNA processes were applied to subsets of EEBO-TCP defined by date-of-printing, comparing works printed between 1520 and 1539 (376 texts, 7.8 million tokens after stop-listing [psl]) with works printed in 1610 and 1611 (274 texts, 6.5 million tokens). These subsets, referred to hereafter as S1520 and S1610, provide snapshots of language use in print for the periods indicated. As will be evident, print output and survival increased significantly between these dates and the chronological parameters were applied to ensure sufficient data for comparison between the two eras. That said, *valour* occurs nearly three times more often in the later dataset and so little in an interim slice of data, that its co-occurrences will not discussed in detail here (see Table 1).

Apart from the difference in raw numbers of occurrences between the two datasets – and noting that the samples are very preliminary snapshots only – what may be learned about *valour* by examining it through the Linguistic DNA windows?

The calculation involved in assessing probability means that the highest PMI scores typically highlight co-occurrences involving a rare term.[12] Thus the highest scoring co-occurring pair-words for *valour* in these two datasets given a reasonable filter level (i.e. a minimum co-occurrence frequency of 5) are *wardeyn* (S1550: 6.861) and *shap* (S1520: 5.977). What we see here is commonly a sign of remaining messy data (one might expect MorphAdorner to standardise *wardeyn*), and/or idiosyncrasies of a particular printed text. For this reason, it can be more informative to take a lemma list formed of terms that have a basic co-occurrence 'above expectation' (operationalised as words with a PMI score of at least 1.0) and order it in terms of these lemmas' own frequency. Combining the outputs from the different window spans (W2, W20, W100 and W200) the first 25 qualifying words from our two main samples are shown in Tables 2a and 2b. As will be observed, while stretching from high to mid-range frequency items in each case, the lists are almost wholly exclusive: *great* is the only high frequency lemma to co-occur noticeably with *valour* in both samples. Expanding to consider all qualifying items from both lists, we find only four other lemmas that appear in both: *courage, loss,*

Table 1. Occurrences and relative frequency of the lemma *valour* in EEBO-TCP works in three samples defined by a specified time span.

Sample	Items printed between	Instances of *valour*	Instances per million words psl
S1520	1520–1539	137	17.56
S1550	1550–1559	69	11.27
S1610	1610–1611	491	75.53

[12]This 'weighting' of low-frequency items is a well-known element of PMI (cf. Manning & Schütze 2001: 182).

Table 2a: First 25 lemmas co-occurring strongly with *valour* in the S1520 dataset when ranked by lemma frequency. Entries shown in italics are excluded by a stoplist in larger window sizes.

		PMI exceeds 1.0 in window of			
Lemma	Frequency in S1520	+/-1	+/-10	+/-50	+/-100
or	*78234*	*			
great	36642	Y	Y		
no	34659	Y	Y		
such	*26224*	*			
before	*21312*	*			
person	7608			Y	
little	7493	Y	Y	Y	
ought	6007		Y		
praise	4992		Y		
wit	4200		Y		
gift	2899			Y	
small	2672		Y	Y	Y
nought	2549			Y	Y
image	2312			Y	**Y**
beside	2307			Y	
clear	2253			Y	
hell	2191			Y	
lack	1952			Y	
forgive	1835			Y	Y
care	1821			Y	
worldly	1725			Y	
sell	1698			Y	Y
just	1639		Y	Y	Y
fore	1623			Y	Y
honest	1538				Y

Table 2b: First 25 lemmas co-occurring strongly with *valour* in the S1610 dataset when ranked by lemma frequency. Entries shown in italics are excluded by a stoplist in larger window sizes.

		PMI exceeds 1.0 in window of			
Lemma	Frequency in S1520	+/-1	+/-10	+/-50	+/-100
his	*116840*	*			
their	*62227*	*			
who	*44501*	*			
my	*29359*	*			
great	23996	Y			
your	*18591*	*			
high	5005		Y		
enemy	4048		Y		
virtue	3715		Y	Y	
war	3558		Y	Y	Y
worthy	3456		Y	Y	
arm	2970		Y	Y	
wisdom	2823		Y		
earl	2495			Y	**Y**
slay	2357			Y	Y
english	2152		Y		
fight	2125		Y	Y	Y
noble	2017			Y	
sword	1952			Y	
field	1747			Y	Y
britain	1688		Y	Y	**Y**
soldier	1539		Y	Y	Y
battle	1468			Y	Y
proof	1453		Y		
edward	1448				Y

princess and *valour* itself. This very limited overlap suggests that *valour* tends to be employed in very different discourses across the two datasets. This may relate to the semantic change observed by Hine. Indeed, we might hypothesise that the co-occurring lexis in Table 2b represents to a greater degree than Table 2a the 'imputation of boldness and courage' observed by Hine (2014), via words such as *virtue, worthy* and *noble*. To affirm or refute this hypothesis, close reading is a necessary next step.

Let us examine a brief example of language in use from each period in order to make some additional observations about how the quantitative output of Linguistic DNA relates to the qualitative study of discourse in texts.

Figure 1 contains a 201-word extract from John Speed's *History of Great Britaine* (EEBO-TCP A12738), centred on an instance of *valour*. Highlighted words belong to a lemma which co-occurs strongly with *valour* in the corresponding LDNA dataset (S1610). Highlighting is used to indicate the innermost window in which the PMI score passes 1.0, with dark/red representing W20, mid/orange W100 and light/yellow W200.

Reading this short passage of Speed's text, one learns that valour is a property of the Saxons as they take the place of Romans as Europe's warring imperialists or 'a second triumphant nation' as Speed puts it. What we learn if we continue to inspect this text in relation to the Linguistic DNA data is that Speed's *History* is a major contributor to the picture created by *valour*'s co-occurrences in the chronological slice of EEBO-TCP. For Speed, valour is powerfully associated with military engagement, conquest and resistance. The lemma's association with Britain in the S1610 dataset, more powerful in the outer window (+/-51–100) than the inner (+/-1–50), arises because Speed makes such

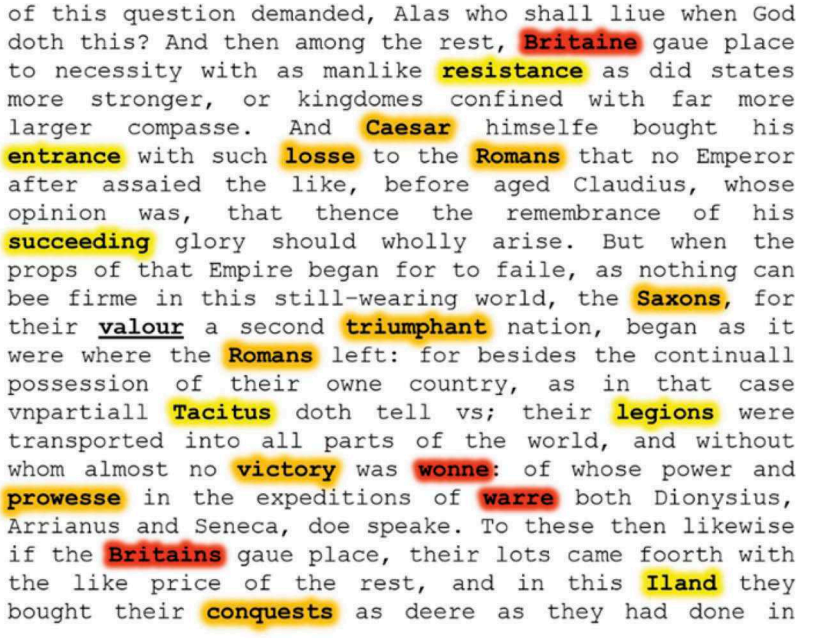
of this question demanded, Alas who shall liue when God doth this? And then among the rest, Britaine gaue place to necessity with as manlike resistance as did states more stronger, or kingdomes confined with far more larger compasse. And Caesar himselfe bought his entrance with such losse to the Romans that no Emperor after assaied the like, before aged Claudius, whose opinion was, that thence the remembrance of his succeeding glory should wholly arise. But when the props of that Empire began for to faile, as nothing can bee firme in this still-wearing world, the Saxons, for their valour a second triumphant nation, began as it were where the Romans left: for besides the continuall possession of their owne country, as in that case vnpartiall Tacitus doth tell vs; their legions were transported into all parts of the world, and without whom almost no victory was wonne: of whose power and prowesse in the expeditions of warre both Dionysius, Arrianus and Seneca, doe speake. To these then likewise if the Britains gaue place, their lots came foorth with the like price of the rest, and in this Iland they bought their conquests as deere as they had done in

Figure 1. An extract from John Speed's *A History of Great Britaine* (London: 1611; EEBO-TCP A12738) highlighted to illustrate instances of *valour*'s strong frequent associates in 1610–1611.

valour a dominant characteristic in his account of British history. In so doing, it would seem, he is appropriating a discourse learned from other European texts, i.e. from the conduct books and, indeed, the chronicles identified by Hine. This is also the text that brings *ann*. and *do*. into the list of *valour*'s associates. Knowing how this one text influences the surrounding dataset suggests caution must be exercised in extrapolating from the basis of the associations mapped using a meagre slice of EEBO-TCP. Nonetheless, the general outcome is encouraging as we look to evolve and refine Linguistic DNA's processes further.

And what of S1520 and the short list of shared lemmas? Figure 2 represents a comparable excerpt from a doctrinal Christian text printed in 1537. Here we see how, even in quasi-metaphorical application, *valour* seems to be closer to modern *value*. Neither the list of lemmas in Table 2a nor this brief discursive passage suggests that the notion of valour in this period wholly lacks implications of physical power or moral worth. However, the body of associations evidenced in the immediate and wider discourse is substantively different from the era when King James' so-called Authorised Version of the Bible appeared.

It is not that the use of *valour* in a common semantic field with *courage* was altogether absent in English but that (if Hine is correct) vernacular translations of European literature radically increased its currency, making this modern sense dominant. It is reasonable to argue that to the extent that early Modern Britain had an encyclopaedic concept of (or associated with) *valour*, it was 'not born English'. Linguistic DNA processes thus point to the existence of differing discourses and discursive concepts in subsets of EEBO-TCP. Close reading – systematic, rigorous, principled, creative and qualitative – facilitates a deeper understanding of those

it is also to be noted, that it is the the wyll of god our father, that we his *sonnes*, and his *children* shulde in this worlde folowe our heed Christe in pacience, and humilite, and that we shulde beare our own crosse, as Christe dyd his. And that we shulde also hate and abhorre all synne, knowynge for suretie, that who so euer dothe not in his herte hate, and abhorre synne, but rather accompteth the breache and violation of goddis commaundement, but as a lyght matter, and of small weight and importaunce: he estemeth not the price and **valour** of this passyon of Christe, accordynge to the dignitie and worthynes therof, but rather semeth to consent, and as moche as in hym is, to go aboute to cause Chryste to be crucified ageyne. In the .v. article it is to be noted, Artycle. Ro. x. that therin is included and conteyned the groundes and foundations of the greattest parte of all the misteries of our catholyque faythe. In so moche that *saynt* Paule sayth, that whosoeuer beleueth in his harte, that god the father dyd resuscitate, and raise vppe his *sonne* Christe from deathe to lyfe, he shall be

Figure 2. An extract from the *Institution of a Christen Man* (London: Thomas Berthelet, 1537; EEBO-TCP A73731) highlighted to illustrate instances of *valour*'s strong frequent associates and with italics for very infrequent co-occurring terms in 1520–1539.

discursive concepts. In this case study, Linguistic DNA processes can be seen as reflecting previous observations on the discourses around *valour*. As the project develops, Linguistic DNA processes will complicate, enhance and perhaps even contradict prior work in a range of disciplines with conceptual interests, including linguistics, literary studies and history.

Finally, we might note that there will always be major and minor notes within any set of associations. The saint, sons and children of the 1537 *Institution* (italicised in Figure 2) represent minor figures in the discourse around *valour*, created at least in part by the presence of that very text in the EEBO-TCP sample. As the project proceeds, we expect that our encyclopaedic endeavours will discover more fully these minor tones too.

8. Conclusion

The significance of Linguistic DNA lies in its combination of ground-breaking computational approaches, a firm theoretical and epistemological foundation in linguistic semantics and pragmatics, and research expertise in philology and historical linguistics. As we bring different areas of interdisciplinary expertise to bear upon one another, we strengthen our shared insights and our capacity for seeing and interpreting the evolution of language and meaning. Such interactions constitute the vanguard of digital humanities and text analytics. The theory, methods and expertise developed within the Linguistic DNA project have application that extends beyond the early modern period, as we continue to refine our processes towards understanding textual meaning in all of its breadth and depth. The definition and operationalisation of *discursive concepts* can be seen as a significant new perspective on what concepts are, and – we anticipate – a valuable complement to existing work in the humanities.

Funding

This work was supported by the Arts and Humanities Research Council [AHRC AH/M00614-X/1].

ORCID

Susan Fitzmaurice http://orcid.org/0000-0002-8804-1987
Justyna A. Robinson http://orcid.org/0000-0001-9392-5720
Marc Alexander http://orcid.org/0000-0002-6337-2632
Iona C. Hine http://orcid.org/0000-0002-9280-5871
Seth Mehl http://orcid.org/0000-0003-3036-8132
Fraser Dallachy http://orcid.org/0000-0002-8694-3565

References

Aarts, Bas, Joanne Close, Geoffrey Leech & Sean Wallis (eds.). 2013. *The verb phrase in English: Investigating recent language change with corpora*. Cambridge: Cambridge University Press.

Allan, Kathryn & Justyna A. Robinson (eds.). 2012. *Current methods in historical semantics*. Berlin: Walter de Gruyter.

Bowie, Jill, Sean Wallis & Bas Aarts. 2013. The perfect in spoken British English. In Bas Aarts, Joanne Close, Geoffrey Leech and Sean Wallis (eds.), 318–352.
Bruni, F. and A. Pettegree (eds.). 2016. *Lost books*. Amsterdam: Brill.
Burns, Philip R. 2013. MorphAdorner v2: A Java library for the morphological adornment of English language texts. Evanston, IL. Northwestern University. https://morphadorner.northwestern.edu/morphadorner/download/morphadorner.pdf. (last accessed on 29 November 2016).
Calzolari, Nicoletta, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odjik, Stelios Piperidis & Daniel Tapias (eds.). 2008. *Proceedings of the Sixth International Language Resources and Evaluation*. Marrakech: European Language Resources Association.
Evans, Vyvyan. 2009. *How words mean: Lexical concepts, cognitive models and meaning construction*. Oxford: Oxford University Press.
Evans, Vyvyan. 2015. A unified account of polysemy within LCCM Theory. *Lingua* 157, 100–123.
Fano, Robert M. 1961. *Transmission of information: A statistical theory of communications*. Boston: MIT Press.
Fitzmaurice, Susan. 2016. Semantic and pragmatic change. In Merja Kytö & Päivi Pahta (eds.), 256–270.
Fitzmaurice, Susan & Irma Taavitsainen (eds.). 2007. *Methods in historical pragmatics*. Berlin: Mouton de Gruyter.
Gadd, Ian. 2009. The use and misuse of *Early English Books Online*. *Literature Compass* 6, 680–692. doi:10.1111/j.1741–4113.2009.00632.x
Geeraerts, Dirk. 1997. *Diachronic prototype semantics: A contribution to historical lexicology*. Oxford: Clarendon Press.
Geeraerts, Dirk. 2010. *Theories of lexical semantics*. Oxford: Oxford University Press.
Geeraerts, Dirk, Stefan Grondelaers & Peter Bakema. 1994. *The structure of lexical variation: Meaning, naming, and context*. Berlin: Mouton de Gruyter.
Geeraerts, Dirk, Caroline Gevaert & Dirk Speelman. 2012. How *anger* rose: Hypothesis testing in diachronic semantics. In Kathryn Allan and Justyna A. Robinson (eds.), 109–132.
Hampsher-Monk, Iain, Karen Tilmans & Frank Van Vree (eds.). 1998. *History of concepts: Comparative perspectives*. Amsterdam: Amsterdam University Press.
Heylen, Kris, Yves Peirsman, Dirk Geeraerts & Dirk Speelman. 2008. Modelling word similarity: An evaluation of automatic synonymy extraction algorithms. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odjik, Stelios Piperidis & Daniel Tapias (eds.), 3243–3249.
Hine, I.C. 2014. Englishing the Bible in early modern Europe: The case of Ruth. Unpublished PhD thesis, University of Sheffield.
Jucker, Andreas H., & Irma Taavitsainen. 2013. English Historical Pragmatics. Edinburgh: Edinburgh University Press.
Kay, Christian, Jane Roberts, Michael Samuels, Irené Wotherspoon & Marc Alexander (eds.). 2016. *The historical thesaurus of English*, version 4.2. Glasgow: University of Glasgow. http://historicalthesaurus.arts.gla.ac.uk (last accessed on 27 January 2017).
Kichuk, Diana. 2007. Metamorphosis: Remediation in *Early English Books Online* (*EEBO*). *Literary and Linguistic Computing* 22(3),291–303. doi: 10.1093/llc/fqm018
Koselleck, Reinhart. 1998. Social history and *Begriffgeschichte*. In Iain Hampsher-Monk, Karen Tilmans and Frank Van Vree (eds.), 23–26.
Koselleck, Reinhart. 2004. *Futures past: On the semantics of historical time*. New York: Columbia University Press.
Kytö, Merja & Päivi Pahta (eds.). 2016. *The Cambridge handbook of English historical linguistics*. Cambridge: Cambridge University Press.
Lehrer, Adrienne. 1992. Names and naming. Why we need fields and frames. In Adrienne Lehrer & Eva F. Kittay (eds.), 123–142.
Lehrer, Adrienne & Eva F. Kittay (eds.). 1992. *Frames, fields and contrasts: New essays in semantic and lexical organization*. Hillsdale, NJ: Erlbaum.

Lenker, U. 2007. *Soþlice, forsooth, truly* – communicative principles and invited inferences in the history of truth-intensifying adverbs in English. In Susan Fitzmaurice and Irma Taavitsainen (eds.), 81–106.
Manning, Christopher & Hinrich Schütze. 2001. *Foundations of statistical natural language processing*. Boston: MIT Press.
Pocock, J. G. A. 1972. *Politics, language and time: Essays on political thought and history*. London: Methuen.
Pocock, J. G. A. 1999–2015. *Barbarism and religion*. 6 volumes. Cambridge: Cambridge University Press.
Porter, Roy. 2001. *Enlightenment: Britain and the creation of the modern world*. Harmondsworth: Penguin.
Roberts, Jane & Christian Kay, with Lynne Grundy. 1995. *A thesaurus of Old English*. King's College London Medieval Studies XI. http://oldenglishthesaurus.arts.gla.ac.uk (last accessed on 27 January 2017).
Samuels, Michael. 1972. *Linguistic evolution, with special reference to English*. Cambridge: Cambridge University Press.
Schama, Simon. 2001. *A history of Britain, volume 2: The British wars, 1603–1776*. London: BBC Worldwide.
Sheskin, David. 2004. *Handbook of parametric and non-parametric statistical procedures*. 3rd edn. Boca Raton: Chapman Hall.
Skinner, Quentin. 1978. *The foundations of modern political thought*. Cambridge: Cambridge University Press.
Smith, Jeremy J. 1996. *An historical study of English: Function, form and change*. London: Routledge.
Taavitsainen, Irma & Susan Fitzmaurice. 2007. Historical pragmatics: What it is and how to do it. In Susan Fitzmaurice and Irma Taavitsainen (eds.), 11–36.
Tucker, Suzie. 1972. *Enthusiasm. A study in semantic change*. Cambridge: Cambridge University Press.
Turney, Peter D. & Patrick Pantel. 2010. From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research* 37, 141–188.
Wallis, S.A. & J. Bowie. 2012. That vexed problem of choice. London: Survey of English Usage. http://www.ucl.ac.uk/english-usage/staff/sean/resources/vexedchoice.pdf (last accessed on 20 January 2017).
Wierzbicka, Anna. 2010. *Experience, evidence, and sense: The hidden cultural legacy of English*. Oxford: Oxford University Press.
Williams, Raymond. 1983. *Keywords: A vocabulary of culture and society*. 2nd edn. Oxford: Oxford University Press.